Inequalities for Bayes Factors and Relative Belief Ratios

by

Zeynep Baskurt
Department of Statistics
University of Toronto

and

Michael Evans
Department of Statistics
University of Toronto

Technical Report No. 1105 September 1, 2011

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Inequalities for Bayes Factors and Relative Belief Ratios

Zeynep Baskurt and Michael Evans
Department of Statistics
University of Toronto

Abstract: We discuss the definition of a Bayes factor, the Savage-Dickey result, and develop some inequalities relevant to Bayesian inferences. We consider the implications of these inequalities for the Bayes factor approach to hypothesis assessment. An approach to hypothesis assessment based on the computation of a Bayes factor, a measure of reliability of the Bayes factor, and the point where the Bayes factor is maximized is recommended. This can be seen to deal with many of the issues and controversies associated with hypothesis assessment. It is noted that an inconsistency in prior assignments can arise when priors are placed on hypotheses that do not arise from a parameter of interest. It is recommended that this inconsistency be avoided by choosing a distance measure from the hypothesis as the parameter of interest. An application is made to assessing the goodness of fit for a logistic regression model and it is shown that this leads to resolving some difficulties associated with assigning priors for this model.

Key words and phrases: Bayes factors, relative belief ratios, inequalities, concentration.

1 Introduction

Bayes factors, as introduced by Jeffreys (1935, 1961), are commonly used in applications of statistics. Kass and Raftery (1995) and Robert, Chopin, and Rousseau (2009) contain detailed discussions of Bayes factors.

Suppose we have a sampling model $\{P_\theta : \theta \in \Theta\}$ on $\mathcal{X}$, and a prior Π on Θ. Let T denote a minimal sufficient statistic for $\{P_\theta : \theta \in \Theta\}$ and $\Pi(\cdot \mid T(x))$ denote the posterior of θ after observing data $x \in \mathcal{X}$. Then for a set $C \subset \Theta$, with $0 < \Pi(C) < 1$, the *Bayes factor in favor of* C is defined by

$$BF(C) = \frac{\Pi(C \mid T(x))}{1 - \Pi(C \mid T(x))} \Big/ \frac{\Pi(C)}{1 - \Pi(C)}.$$

Clearly $BF(C)$ is a measure of how beliefs in the true value being in C have changed from *a priori* to *a posteriori*. Alternatively, we can measure this change

in belief by the *relative belief ratio* of C, namely, $RB(C) = \Pi(C \mid T(x))/\Pi(C)$. A relative belief ratio measures change in belief on the probability scale as opposed to the odds scale for the Bayes factor. While a Bayes factor is the multiplicative factor transforming the prior odds after observing the data, a relative belief ratio is the multiplicative factor transforming the prior probability. These measures are related as we have that

$$BF(C) = \frac{(1-\Pi(C))RB(C)}{1-\Pi(C)RB(C)},\ RB(C) = \frac{BF(C)}{\Pi(C)BF(C)+1-\Pi(C)}, \quad (1)$$

and $BF(C) = RB(C)/RB(C^c)$. If it is hypothesized that $\theta \in H_0 \subset \Theta$, then $BF(H_0)$ or $RB(H_0)$ can be used as an assessment as to what extent the observed data has changed our beliefs in the true value being in H_0.

Both the Bayes factor and the relative belief ratio are not defined when $\Pi(C) = 0$. In Section 2 we will see that, when we have a characteristic of interest $\psi = \Psi(\theta)$, where $\Psi : \Theta \to \Psi$ (we don't distinguish between the function and its range to save notation), and $H_0 = \Psi^{-1}\{\psi_0\}$ with $\Pi(H_0) = 0$, we can define the Bayes factor and relative belief ratio of H_0 as limits and that the limiting values are identical. This permits the assessment of a hypothesis $H_0 = \Psi^{-1}\{\psi_0\}$ via a Bayes factor without the need for placing prior mass on ψ_0.

In a number of circumstances there is a set $H_0 \subset \Theta$, with $\Pi(H_0) = 0$, that we want to assess and there is no particular characteristic of interest $\psi = \Psi(\theta)$ for which $H_0 = \Psi^{-1}\{\psi_0\}$. For example, this arises in the Behrens-Fisher problem and in factor analysis. A common approach here is to use a mixture prior $\Pi_\gamma = \gamma\Pi_0 + (1-\gamma)\Pi$ where $\gamma \in (0,1)$ and Π_0 is a prior on H_0. In Section 3 we point to a general inconsistency in this approach and how this can be resolved. In Section 5 we propose the method of concentration as a general resolution of this problem and address some computational issues. In Section 6 we apply this to obtain a Bayesian goodness-of-fit test for the logistic regression model and show that this leads to priors similar to those recommended by Bedrick, Christensen and Johnson (1996, 1997) but avoiding an ambiguity in that approach.

One concern with both Bayes factors and relative belief ratios is how they should be calibrated. From (1) we see that, for fixed $\Pi(C)$, $BF(C)$ is an increasing function of $RB(C)$ and conversely, so it is somewhat of a personal choice as to which to use to measure change in belief. We will discuss the calibration of these quantities in Section 4 together with some relevant inequalities. Also we establish some close parallels between use of Bayes factors to assess statistical evidence and the approach to assessing statistical evidence via likelihood ratios discussed in Royall (1997, 2000).

More general definitions have been offered for Bayes factors when improper priors are employed. O'Hagan (1995) defines fractional Bayes factors and Berger and Perrichi (1996) define intrinsic Bayes factors. In this paper we restrict attention to proper priors although limiting results can often be obtained when considering a sequence of increasingly diffuse priors. Lavine and Schervish (1999) consider the coherency behavior of Bayes factors.

2 Definition for a Marginal Parameter

We now extend the definition of relative belief ratios and Bayes factors to the case where $\Pi(H_0) = 0$. We assume that P_θ has density f_θ with respect to support measure μ, Π has density π on Θ with respect to support measure ν and $\pi(\cdot \mid T(x))$ denotes the posterior density on Θ with respect to ν. Suppose we wish to assess $H_0 = \Psi^{-1}\{\psi_0\}$ for some parameter of interest $\psi = \Psi(\theta)$.

We will suppose that all our spaces possess sufficient structure, and the various mappings we consider are sufficiently smooth, so that the support measures are volume measure on the respective spaces and that any densities used are derived as limits of the ratios of measures of sets converging to points (see, for example, Rudin (1974), Chapter 8). In particular, whenever a density exists which is continuous at a point this limiting process produces that value for the density and note that this is necessary for the interpretation of continuous probability as an approximation to discrete models, i.e., densities cannot be arbitrarily defined at such points. We do not go into further details on this point here except to say that these restrictions are typically satisfied in statistical problems. The mathematical details can be found in Tjur (1974), where it is seen that we effectively require Riemann manifold structure for the various spaces considered. For example, these requirements are always satisfied in the discrete case, as well as in the case of the commonly considered continuous statistical models. One appealing consequence of such restrictions is that we get simple formulas for marginal and conditional densities. For example, putting $J_\Psi(\theta) = (\det(d\Psi(\theta))(d\Psi(\theta))^t)^{-1/2}$ where $d\Psi$ is the differential of Ψ, and supposing $J_\Psi(\theta)$ is finite and positive for all θ, then the prior probability measure Π_Ψ has density, with respect to volume measure ν_Ψ on Ψ, given by

$$\pi_\Psi(\psi) = \int_{\Psi^{-1}\{\psi\}} \pi(\theta) J_\Psi(\theta)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta), \tag{2}$$

where $\nu_{\Psi^{-1}\{\psi\}}$ is volume measure on $\Psi^{-1}\{\psi\}$. Furthermore, the conditional prior density of θ given $\Psi(\theta) = \psi$ is

$$\pi(\theta \mid \psi) = \pi(\theta) J_\Psi(\theta) / \pi_\Psi(\psi) \tag{3}$$

with respect to $\nu_{\Psi^{-1}\{\psi\}}$ on $\Psi^{-1}\{\psi\}$. A significant advantage with (2) and (3) is that there is no need to introduce coordinates, as is commonly done, for so-called nuisance parameters. In general, such coordinates do not exist.

If we let $T : \mathcal{X} \to \mathcal{T}$ denote a minimal sufficient statistic for $\{f_\theta : \theta \in \Theta\}$, then the density of T, with respect to volume measure $\mu_\mathcal{T}$ on $\mathcal{T}$, is given by $f_{\theta T}(t) = \int_{T^{-1}\{t\}} f_\theta(x) J_T(x)\, \mu_{T^{-1}\{t\}}(dx)$, where $\mu_{T^{-1}\{t\}}$ denotes volume on $T^{-1}\{t\}$. The prior predictive density, with respect to μ, of the full data is given by $m(x) = \int_\Theta \pi(\theta) f_\theta(x)\, \nu(d\theta)$ and the prior predictive density of T, with respect to $\mu_\mathcal{T}$, is then $m_T(t) = \int_\Theta \pi(\theta) f_{\theta T}(t)\, \nu(d\theta) = \int_{T^{-1}\{t\}} m(x) J_T(x)\, \mu_{T^{-1}\{t\}}(dx)$. This leads to a generalization of the Savage-Dickey ratio result, see Dickey and Lientz (1970), Dickey (1971), as we don't require coordinates for nuisance parameters.

Theorem 1. (*Savage-Dickey*) $\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi) = m_T(T(x) \,|\, \psi)/m_T(T(x))$.
Proof: The posterior density of θ is $\pi(\theta \,|\, T(x)) = \pi(\theta) f_{\theta T}(T(x))/m_T(T(x))$, and the posterior density of $\psi = \Psi(\theta)$, with respect to ν_Ψ, is $\pi_\Psi(\psi \,|\, T(x)) = \int_{\Psi^{-1}\{\psi\}} (\pi(\theta) f_{\theta T}(T(x))/m_T(T(x))) J_\Psi(\theta)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta) = \pi_\Psi(\psi) \int_{\Psi^{-1}\{\psi\}} \pi(\theta \,|\, \psi)$ $(f_{\theta T}(T(x))/m_T(T(x)))\, \nu_{\Psi^{-1}\{\psi\}}(d\theta) = \pi_\Psi(\psi) m_T(T(x) \,|\, \psi)/m_T(T(x))$ where $m_T(\cdot \,|\, \psi)$ is the conditional prior predictive density of T, given $\Psi(\theta) = \psi$.
As T is minimal sufficient, $m_T(T(x) \,|\, \psi)/m_T(T(x)) = m(x \,|\, \psi)/m(x)$.

Since $\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi)$ is the density of $\Pi_\Psi(\cdot \,|\, T(x))$ with respect to Π_Ψ, we have that

$$\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi) = \lim_{\epsilon \to 0} \Pi_\Psi(C_\epsilon(\psi) \,|\, T(x))/\Pi_\Psi(C_\epsilon(\psi)) \tag{4}$$

whenever $C_\epsilon(\psi)$ converges nicely (see Rudin (1974) for the definition of 'nicely') to $\{\psi\}$ as $\epsilon \to 0$ and all densities are continuous at ψ, e.g., $C_\epsilon(\psi)$ could be a ball of radius ϵ centered at ψ. So $\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi)$ is the limit of the relative belief ratios of sets converging to ψ and, if $\Pi(\Psi^{-1}\{\psi\}) > 0$, then $\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi)$ gives the previous definition of a relative belief ratio for $\Psi^{-1}\{\psi\}$. As such we will refer to $RB(\psi) = \pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi)$ as the *relative belief ratio* of ψ.

From (4) and (1) we have that $BF(C_\epsilon(\psi)) \to (1 - \Pi(\Psi^{-1}\{\psi\}))RB(\psi)/(1 - \Pi(\Psi^{-1}\{\psi\})RB(\psi))$ as $\epsilon \to 0$ and this equals $RB(\psi)$ if and only if $\Pi(\Psi^{-1}\{\psi\}) = 0$. So, in the continuous case, $RB(\psi)$ is a limit of Bayes factors with respect to Π and so can also be called the Bayes factor in favor of ψ with respect to Π. If, however, $\Pi(\Psi^{-1}\{\psi\}) > 0$, then $RB(\psi)$ is not a Bayes factor with respect to Π but is related to the Bayes factor through (1). We see that, under general conditions, whenever $H_0 = \Psi^{-1}\{\psi_0\}$ we can assess this hypothesis via the relative belief ratio $RB(\psi_0)$, that this is a limit of Bayes factor in the continuous case, and is closely related to the Bayes factor generally.

3 A General Hypothesis

Consider a null hypothesis $H_0 \subset \Theta$ and suppose we have not specified a smooth Ψ of interest and a value ψ_0 so that Ψ and ψ_0 *generate* H_0 via $H_0 = \Psi^{-1}\{\psi_0\}$. When $\Pi(H_0) > 0$, then $\Psi = I_{H_0}$ is smooth in the discrete topology on $\{0, 1\}$ and $H_0 = \Psi^{-1}\{1\}$. While there is nothing that forces us to make this choice of Ψ, both the relative belief ratio and Bayes factor of H_0 are, in this case, invariant to the choice of Ψ. This is not the case, however, when $\Pi(H_0) = 0$, as will happen whenever Π is dominated by volume measure and H_0 is a lower dimensional subset of Θ. It is then not clear how to compute the relative belief ratio or Bayes factor for H_0.

The usual approach when $\Pi(H_0) = 0$ is to assign a prior probability $0 < \gamma < 1$ to H_0 and replace Π by $\Pi_\gamma = \gamma\Pi_0 + (1 - \gamma)\Pi$ where Π_0 is a probability measure on H_0. Since $\Pi_\gamma(H_0) = \gamma$, we obtain the Bayes factor and relative belief ratio for H_0 with respect to Π_γ as $m_{0T}(T(x))/m_T(T(x))$ and $\{m_{0T}(T(x))/m_T(T(x))\}/\{1 - \gamma + \gamma m_{0T}(T(x))/m_T(T(x))\}$, respectively, where

$m_{0T}(T(x)) = \int_{H_0} f_{\theta T}(T(x))\,\Pi_0(d\theta)$. Note that these expressions become equal as $\gamma \to 0$ and that these expressions require $\Pi(H_0) = 0, \Pi_0(H_0^c) = 0$.

Now suppose we consider a Ψ and ψ_0 that generates H_0. We then have the following result generalizing Verdinelli and Wasserman (1995) as we don't require coordinates for nuisance parameters.

Theorem 2. (*Verdinelli-Wasserman*) When $H_0 = \Psi^{-1}\{\psi_0\}$ for some Ψ and ψ_0 and $\Pi(H_0) = 0$, then the Bayes factor in favor of H_0 with respect to Π_γ is

$$m_{0T}(T(x))/m_T(T(x)) = RB(\psi_0)E_{\Pi_0}\left(\pi(\theta \mid \psi_0, T(x))/\pi(\theta \mid \psi_0)\right) \tag{5}$$

where E_{Π_0} refers to expectation with respect to Π_0.
Proof: Since $m_{0T}(T(x))/m_T(T(x)) = RB(\psi_0)m_{0T}(T(x))/m_T(T(x) \mid \psi_0)$ and

$$\frac{m_{0T}(T(x))}{m_T(T(x) \mid \psi_0)} = \frac{\int_{\Psi^{-1}\{\psi_0\}} \pi_0(\theta) f_{\theta T}(T(x))\,\nu_{\Psi^{-1}\{\psi_0\}}(d\theta)}{\int_{\Psi^{-1}\{\psi_0\}} \pi(\theta \mid \psi_0) f_{\theta T}(T(x))\,\nu_{\Psi^{-1}\{\psi_0\}}(d\theta)},$$

the result follows from (3).

When $\Pi_0 = \Pi(\cdot \mid \psi_0)$, then $E_{\Pi_0}\left(\pi(\theta \mid \psi_0, T(x))/\pi(\theta \mid \psi_0)\right) = 1$ and (5) establishes that $RB(\psi_0)$ is a Bayes factor with respect to Π_γ. We see from Section 2, however, that we really do not have to introduce the measure Π_γ to interpret $RB(\psi_0)$ as a Bayes factor, as it is a limit of Bayes factors with respect to Π.

In general (5) establishes the relationship between the Bayes factor when using the conditional prior $\Pi(\cdot \mid \psi_0)$ on H_0 and the Bayes factor when using the prior Π_0 on H_0. The adjustment is the expected value, with respect to Π_0, of the conditional relative belief ratio $\pi(\theta \mid \psi_0, T(x))/\pi(\theta \mid \psi_0)$ for $\theta \in H_0$, given H_0. This can also be written as $E_{\Pi(\cdot \mid \psi_0, T(x))}\left(\pi_0(\theta)/\pi(\theta \mid \psi_0)\right)$ and so measures the discrepancy between the conditional priors given H_0 under Π and Π_γ. So when π_0 is substantially different than $\pi(\cdot \mid \psi_0)$ we can expect a significant difference in the Bayes factors. Clearly, there is an inconsistency in the prior assignments if Π_0 is not equal to $\Pi(\cdot \mid \psi_0)$ for some smooth Ψ and ψ_0.

This inconsistency can also be seen at a fundamental level when we consider how the priors are expressing beliefs about θ given that H_0 is true. For our conditional relative prior belief for the values $\theta_1, \theta_2 \in H_0$ is given by $\pi(\theta_1)J_\Psi(\theta_1)/\pi(\theta_2)J_\Psi(\theta_2)$ when using $\Pi(\cdot \mid \psi_0)$ on H_0 and by $\pi_0(\theta_1)/\pi_0(\theta_2)$ when using Π_0 on H_0. Unless these ratios are equal for some Ψ there is an inconsistency expressed in our prior beliefs between the priors Π and Π_γ. So a natural requirement for any measure Π_0 is that $\pi_0(\theta_1)/\pi_0(\theta_2) = \pi(\theta_1)J_\Psi(\theta_1)/\pi(\theta_2)J_\Psi(\theta_2)$ for some Ψ, for all $\theta_1, \theta_2 \in H_0$. When we do this there is no need to introduce the prior Π_γ as we can proceed via Theorem 1 and treat $RB(\psi_0)$ as the relevant Bayes factor and relative belief ratio. If one doubts the need for this consistency requirement, consider the discrete case or the case when we have a Ψ of interest, as it seems clear then that choosing a Π_0 that has no relation to Π is incorrect.

The question remains, however, which of the many Ψ functions that generate H_0 do we use when there is no particular parameter of interest? Note that this problem is similar to the well-known Borel paradox of probability theory, as it is not a well-posed problem, and as such does not have an ultimate solution. It

still seems more appropriate, however, to choose a Ψ rather than a Π_0 as, in the former case, we are guaranteed consistency of prior assignments on H_0. Before proposing a general answer to this question we consider a simple approach that is sometimes available. Suppose Ψ is smooth, $H_0 = \Psi^{-1}\{\psi_0\}$ and $J_\Psi(\theta)$ is constant and positive for $\theta \in H_0$. We will refer to such a Ψ as a *constant volume distortion generator of* H_0 and denote the class of such transformations by $\mathcal{T}_{H_0}$. We have the following result.

Theorem 3. For every $\Psi \in \mathcal{T}_{H_0}$ the conditional relative prior belief of $\theta_1, \theta_2 \in H_0$ is given by $\pi(\theta_1)/\pi(\theta_2)$. Furthermore, the relative belief ratio of H_0 is independent of $\Psi \in \mathcal{T}_{H_0}$.

Proof: For $\theta_1, \theta_2 \in H_0$ we have that $\pi(\theta_1)J_\Psi(\theta_1)/\pi(\theta_2)J_\Psi(\theta_2) = \pi(\theta_1)/\pi(\theta_2)$ which proves the first part. Also $\pi_\Psi(\psi_0) = \int_{\Psi^{-1}\{\psi_0\}} \pi(\theta)J_\Psi(\theta)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta)$ and $\pi_\Psi(\psi_0 \,|\, T(x)) = \int_{\Psi^{-1}\{\psi_0\}} \pi(\theta \,|\, T(x))J_\Psi(\theta)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta)$ and since $J_\Psi(\theta)$ is constant inside the integrals, this proves the second statement.

Note that Theorem 3 generalizes the invariance of the Bayes factor and relative belief ratio to Ψ found in the discrete case, to the situation $\Pi(H_0) = 0$ and $\Psi \in \mathcal{T}_{H_0}$. In essence, when $\Psi \in \mathcal{T}_{H_0}$, the volume distortions induced by Ψ, as measured by $J_\Psi(\theta)$, do not affect our conditional prior beliefs about $\theta \in H_0$ and these are essentially given by the values of $\pi(\theta)$ for $\theta \in H_0$. These results provide an argument for basing the computation of a Bayes factor (or relative belief ratio) for H_0 on a $\Psi \in \mathcal{T}_{H_0}$, when such a transformation is available.

We consider some examples.

Example 1. *Point null hypothesis in* Θ.

Suppose that $H_0 = \{\theta_0\}$ where $\Pi(H_0) = 0$. Then the only possible probability measure on H_0 is degenerate at θ_0 and indeed $\Pi_0 = \Pi(\cdot \,|\, \theta_0)$. Therefore, using Theorem 2, we have that the Bayes factors for H_0 under Π and Π_γ are the same. Again we note that there is no need to introduce the mixture prior Π_γ as the Bayes factor $f_{\theta_0 T}(T(x))/m_T(T(x))$ arises from Π as a limit. Also, any smooth transformation on Θ is in $\mathcal{T}_{H_0}$ and so Theorem 3 applies.

Example 2. *Behrens-Fisher problem.*

Suppose that x is a sample from a $N(\mu_1, \sigma_1^2)$ distribution independent of a sample y from a $N(\mu_2, \sigma_2^2)$ distribution where $(\mu_i, \sigma_i^2) \in R \times R^+$ are unknown for $i = 1, 2$ and we wish to assess the hypothesis $H_0 : \mu_1 = \mu_2$. There are a number of different possibilities for Ψ but perhaps the simplest is $\psi = \Psi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \mu_1 - \mu_2$ as $H_0 = \Psi^{-1}\{0\}$. This choice of Ψ has $J_\Psi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = 1/\sqrt{2}$ and so satisfies Theorem 3. There are many other elements of $\mathcal{T}_{H_0}$, for example, any 1-1 smooth function of Ψ is in $\mathcal{T}_{H_0}$, but they all lead to the same value of the Bayes factor. There are many other Ψ transformations that generate $H_0 : \mu_1 = \mu_2$. For example, for $p > 0$, using $\psi = \Psi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \mu_1^p - \mu_2^p$ we have $H_0 = \Psi^{-1}\{0\}$. Then $J_\Psi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = p^{-1}(\mu_1^{2(p-1)} - \mu_2^{2(p-1)})^{-1/2}$ which equals $p^{-1}\mu^{-(p-1)}$ when $\mu_1 = \mu_2 = \mu$ and this is not constant on H_0 when $p \neq 1$.

Example 3. *Regression.*

Suppose we have $y = X\beta + \sigma e$ where $X \in R^{n\times k}$ is known and of rank $k, \beta \in R^k$ and $\sigma > 0$ are unknown, e is a sample from a known distribution with mean 0 and variance 1, and we want to assess $H_0 : X\beta \in \mathcal{L}$ where $\mathcal{L}$ is a linear subspace of R^n of dimension $l < k$. Let $L \in R$ be such that the columns of L form a basis of $\mathcal{L}$, let $A = (X'X)^{-1}X'L \in R^{k\times l}$ and $B \in R^{k\times(k-l)}$ be such that (A, B) has rank k and $B'A = 0$. Then H_0 is true if and only if $B'\beta = 0$. Letting $\Psi(\beta, \sigma) = B'\beta$ we have that $H_0 = \Psi^{-1}\{0\}$ and $J_\Psi(\beta, \sigma) = (\det(B'B))^{-1/2}$ so $\Psi \in \mathcal{T}_{H_0}$.

In general $\mathcal{T}_{H_0}$ could be empty. To ensure the consistency of prior assignments it is necessary, however, to select a Ψ rather than Π_0. In Section 5 we discuss a general approach to selecting Ψ that has good intuitive support.

4 Calibration and Inequalities

A Bayes factor or relative belief ratio for H_0 measures how our beliefs in H_0 change after seeing the data. It is relevant to ask, however, how reliable is this evidence? One way to answer this is to propose a scale on which a Bayes factor can be assessed. For example, Kass and Raftery (1995) discuss using a scale due to Jeffreys (1961). Comparing the Bayes factor to such a scale, however, does not answer a very natural question: how well do the data support alternatives to H_0? For example, when $H_0 = \Psi^{-1}\{\psi_0\}$ we can consider the Bayes factor for other values of ψ. If a Bayes factor for a $\psi \neq \psi_0$ is much larger than that for ψ_0, then it seems reasonable to at least express some doubt as to the reliability of the evidence in favour of H_0. Note that we are proposing to compare $RB(\psi_0)$ to each of the possible values of $RB(\psi)$ as part of assessing H_0, as opposed to just considering the hypothesis testing problem H_0 versus H_0^c (see, however, Example 5). This is in agreement with a commonly held view as expressed, for example, in Gelman, Carlin, Stern and Rubin (2004) concerning hypothesis assessment. Note that this is different than, for example, the discussion in Berger and Delampady (1987).

We consider first measuring the reliability of a relative belief ratio and recall that this is also a Bayes factor in the continuous case. Perhaps the most obvious way to do this is via the tail probability

$$\Pi_\Psi\left(RB(\psi) \leq RB(\psi_0) \mid T(x)\right). \tag{6}$$

This is the posterior probability of a value ψ having a relative belief no greater than $RB(\psi_0)$. If this probability is small, then there are other values of ψ with greater relative beliefs and the data leads us to assign a large amount of our belief to such values and this certainly should lead to some doubt concerning H_0. Note that we are not suggesting that a small value of (6) necessarily be interpreted as evidence against H_0. A small value of (6) does suggest, however, that any evidence in favour of H_0, as expressed via $RB(\psi_0)$, is not very reliable. Interpreting the value of a Bayes factor or relative belief ratio without some assessment of the uncertainty does not seem statistically appropriate.

The following simple inequality holds.

Theorem 4. The value of (6) is bounded above by $RB(\psi_0)$, namely,

$$\Pi_\Psi\left(RB(\psi) \le RB(\psi_0) \mid T(x)\right) \le RB(\psi_0). \tag{7}$$

Proof: $\Pi_\Psi\left(RB(\psi) \le RB(\psi_0) \mid T(x)\right) = \int_{\{\psi: RB(\psi) \le RB(\psi_0)\}} \pi_\Psi(\psi \mid T(x))\, \nu_\Psi(d\psi)$ $\le \int_{\{\psi: RB(\psi) \le RB(\psi_0)\}} RB(\psi_0)\pi_\Psi(\psi)\, \nu_\Psi(d\psi) \le RB(\psi_0)$.

Of course (7) tells us nothing when $RB(\psi_0) \ge 1$. It does tell us, however, that a very small value of $RB(\psi_0)$ is very strong and reliable evidence against H_0 and in fact there is no need to compute (6) in such a situation.

When $\Pi(\Psi^{-1}\{\psi\}) = 0$ we can also interpret $RB(\psi_0)$ as the Bayes factor with respect to Π in favour of H_0 and so (6) is also a calibration of the Bayes factor. When ψ has a discrete distribution, we have the following result where we interpret $BF(\psi)$ in the obvious way.

Corollary 5. If Π_Ψ is discrete, then $\Pi_\Psi\left(BF(\psi) \le BF(\psi_0) \mid T(x)\right) \le BF(\psi_0) \times$ $E_\Pi(\{1 + \pi_\Psi(\Psi(\theta))(BF(\psi_0) - 1)\}^{-1})$, the upper bound is finite and converges to 0 as $BF(\psi_0) \to 0$.

Proof: Using (1) we have that $BF(\psi) \le BF(\psi_0)$ if and only if $RB(\psi) \le BF(\psi_0)/\{1 + \pi_\Psi(\psi)(BF(\psi_0) - 1)\}$ and, as in the proof of Theorem 4, this implies the inequality. Also $1 + \pi_\Psi(\psi)(BF(\psi_0) - 1) \ge 1 + \max_\psi \pi_\Psi(\psi)(BF(\psi_0) - 1)$ when $BF(\psi_0) \le 1$ and $1 + \pi_\Psi(\psi)(BF(\psi_0) - 1) \ge 1 + \min_\psi \pi_\Psi(\psi)(BF(\psi_0) - 1)$ when $BF(\psi_0) > 1$ which completes the proof.

So we see that a small value of $BF(\psi_0)$ is, in both the discrete and continuous case, reliable evidence against H_0.

It is natural to ask if large values of $RB(\psi_0)$ and $BF(\psi_0)$ are always reliable evidence in favour of H_0? Consider the following example.

Example 4. *Location normal.*

Suppose we have a sample $x = (x_1, \ldots, x_n)$ from a $N(\mu, 1)$ distribution, where $\mu \in R^1$ is unknown, so $T(x) = \bar{x}$, we take $\mu \sim N(0, \tau^2), \Psi(\mu) = \mu$, and we want to assess $H_0 : \mu = 0$. In Figure 1 we have plotted $RB(0)$ and (6) against $\sqrt{n}\bar{x}$ for several cases. From this we can see that, for a given value of $\sqrt{n}\bar{x}$, there is a huge discrepancy between the upper bound and (6) as either n or τ^2 grows. If we accept (6) as the appropriate calibration of the ratio $RB(0)$, then there is clearly a problem with reliably interpreting large values of $RB(0)$ as evidence in support of H_0.

Based on Jeffreys' scale $RB(0) = 20.72$ is strong evidence in favour of H_0, but when $n = 50, \tau^2 = 400$, then (6) equals 0.05 and, as such, 20.72 does not seem to be overwhelming evidence in favour of H_0. The value $\hat{\mu}$ maximizing $RB(\mu)$, see Figure 2, is given by $\hat{\mu} = 0.28$ and $RB(\hat{\mu}) = 141.40$. Note that $\hat{\mu} = 0.28$ cannot be interpreted as being close to 0 independent of the application context. If, however, the application dictates that a value of 0.28 is practically speaking close enough to 0 to be treated as 0, then it certainly seems reasonable to proceed as if H_0 is correct and this is supported by the value of the Bayes factor. In general, it can be shown that for a fixed value of $RB(0)$, then (6) decreases to 0 as either n or τ^2 grows. Basically this is saying that a higher standard is set

for saying a fixed value of $RB(0)$ is evidence in favour of H_0, as we increase the amount of data or make the prior more diffuse. So, for example, the more data we have, the larger $RB(0)$ has to be convincing evidence in favour of H_0.

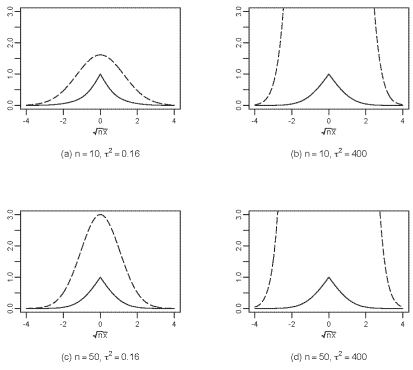

Figure 1: Plot of $\Pi_\Psi\left(RB(\psi) \leq RB(\psi_0) \,|\, T(x)\right)$ (-) and $RB(\psi_0)$ (- -) against $\sqrt{n}\bar{x}$ for various choices of n and τ^2 in Example 4.

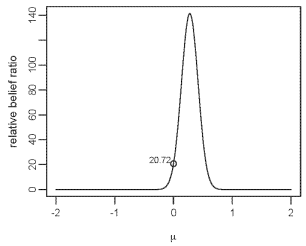

Figure 2: Plot of $RB(\mu)$ against μ when $n = 50, \tau^2 = 400$ and $\sqrt{n}\bar{x} = 1.96$ in Example 4.

The following example is concerned with comparing H_0 to H_0^c.

Example 5. *Binary* Ψ.

Suppose $\Psi(\theta) = I_{H_0}$ and $0 < \Pi(H_0) < 1$. So $\Pi_\Psi\left(BF(\psi) \leq BF(H_0) \,|\, T(x)\right) = \Pi(H_0 \,|\, T(x))$ when $BF(H_0) \leq 1$, and $\Pi_\Psi\left(BF(\psi) \leq BF(H_0) \,|\, T(x)\right) = 1$ otherwise, while $\Pi_\Psi\left(RB(\psi) \leq RB(H_0) \,|\, T(x)\right) = \Pi(H_0 \,|\, T(x))$ when $BF(H_0) \leq 1$, and $\Pi_\Psi\left(RB(\psi) \leq RB(H_0) \,|\, T(x)\right) = 1$ otherwise. So these give the same assessment of reliability. This says that in the binary case $BF(H_0) < 1$ or $RB(H_0) < 1$ is reliable evidence against H_0 only when $\Pi(H_0 \,|\, T(x))$ is small. By Corollary 5 and Theorem 4 this will be the case whenever $BF(H_0)$ or $RB(H_0)$ are suitably small. Furthermore, large values of $BF(H_0)$ or $RB(H_0)$ are always deemed to be reliable evidence in favour of H_0 in this case. Note that $BF(H_0) = m_{H_0}(T(x))/m_{H_0^c}(T(x))$ where $m_{H_0}, m_{H_0^c}$ are the prior predictive densities of T under H_0 and H_0^c, respectively.

So if one has determined in an application that comparing H_0 to H_0^c is the appropriate approach, as opposed to comparing the hypothesized value of the parameter of interest to each of its alternative values, then (6) leads to the usual

answers. We might refer to this as hypothesis testing as opposed to hypothesis assessment. We advocate hypothesis assessment, however, particularly as it allows us to avoid the inconsistency in prior assignments discussed in Section 3.

Using Theorem 1 we can express Theorem 4 as follows.

Corollary 6. $\Pi_\Psi\left(m_T(T(x)\,|\,\psi)/m_T(T(x)) \leq RB(\psi_0)\,|\,T(x)\right) \leq RB(\psi_0)$.

Note that $m_T(T(x)\,|\,\psi_0)/m_T(T(x))$ can be interpreted as the relative belief in the value $T(x)$ from *a priori* Π to *a priori* $\Pi(\cdot\,|\,\psi_0)$. We also have a coarser bound on (6).

Corollary 7. $\Pi_\Psi\left(RB(\psi) \leq RB(\psi_0)\,|\,T(x)\right) \leq \sup_{\theta\in\Psi^{-1}\{\psi_0\}} f_{\theta T}(t)/m_T(T(x))$.

Proof: We have that $m_T(T(x)\,|\,\psi_0) = \int_{\Psi^{-1}\{\psi_0\}} \pi(\theta\,|\,\psi_0) f_{\theta T}(t)\,\nu_{\Psi^{-1}\{\psi_0\}}(d\theta) \leq \sup_{\theta\in\Psi^{-1}\{\psi_0\}} f_{\theta T}(t)$ and the result follows by Theorem 1.

We see that $\sup_{\theta\in\Psi^{-1}\{\psi_0\}} f_{\theta T}(t)/m_T(T(x))$ is a standardized profile likelihood at ψ_0, i.e., it cannot be multiplied by a positive constant. So the profile likelihood has an evidential interpretation as part of an upper bound on (6). The proof of Theorem 1 shows $RB(\psi_0) = \int_{\Psi^{-1}\{\psi\}} \pi(\theta\,|\,\psi)(f_{\theta T}(T(x))/m_T(T(x)))\,\nu_{\Psi^{-1}\{\psi\}}(d\theta)$ is a standardized integrated likelihood at ψ_0 and it gives a sharper bound on (6). This can be interpreted as saying the integrated likelihood contains more relevant information concerning H_0 than the profile likelihood. This provides support for the use of integrated likelihoods over profile likelihoods as discussed in Berger, Liseo, and Wolpert (1999). If we simply treat $RB(\psi)$ as an integrated likelihood, however, then we lose the interpretation of the ratio as a relative belief and we lose (7).

Now suppose we follow Royall (2000) and consider the prior probability of getting a small value of $RB(\psi_0)$ when H_0 is true, as we know from (7) that this would be misleading evidence. We have the following result.

Theorem 8. The prior probability that $RB(\psi_0) \leq 1/k$, given that H_0 is true, is bounded above by $1/k$.

Proof: We have that

$$\Pi\times P_\theta\left(\frac{\pi_\Psi(\psi_0\,|\,T(X))}{\pi_\Psi(\psi_0)} \leq \frac{1}{k}\,\middle|\,\psi_0\right) = \Pi\times P_\theta\left(\frac{m_T(T(X)\,|\,\psi_0)}{m_T(T(X))} \leq \frac{1}{k}\,\middle|\,\psi_0\right)$$
$$= \int_{\left\{t:\frac{m_T(t\,|\,\psi_0)}{m_T(t)}\leq\frac{1}{k}\right\}} m_T(t\,|\,\psi_0)\,\mu_T(dt) \leq \int_{\left\{t:\frac{m_T(t\,|\,\psi_0)}{m_T(t)}\leq\frac{1}{k}\right\}} \frac{m_T(t)}{k}\,\mu_T(dt) \leq \frac{1}{k}.$$

Theorem 8 tells us that, *a priori*, the relative belief ratio for H_0 is unlikely to be small when H_0 is true.

The probability in Theorem 8 equals $M_T\left(m_T(t\,|\,\psi_0)/m_T(t) \leq 1/k\,|\,\psi_0\right)$. Then choosing $k = m_T(T(x))/m_T(T(x)\,|\,\psi_0)$ gives the following result.

Corollary 9. $M_T\left(m_T(t\,|\,\psi_0)/m_T(t) \leq RB(\psi_0)\,|\,\psi_0\right) \leq RB(\psi_0)$.

We can interpret the left-hand side of the inequality as a frequentist tail probability for assessing H_0. If we consider $\theta \sim \Pi(\cdot\,|\,\psi_0)$ and $x \sim P_\theta$, then the left-hand side is the prior probability, when H_0 is true, of obtaining data that would lead to a relative belief ratio no greater than that observed. On the other hand (6) is a Bayesian calibration of $RB(\psi_0)$. We see that the ratio $RB(\psi_0)$

provides an upper bound on both of these tail probabilities and so a small value of $RB(\psi_0)$ leads to evidence against H_0 from both points of view. A similar conclusion holds for the upper bound $\sup_{\theta\in\Psi^{-1}\{\psi_0\}} f_{\theta T}(t)/m_T(T(x))$.

Theorem 8 is concerned with $RB(\psi_0)$ providing misleading information when H_0 is true. Naturally, we are also concerned with the prior probability that $RB(\psi)$ is large when H_0 is false, namely, when $\psi = \Psi(\theta) \neq \psi_0$. For this we consider the behavior of the ratio $RB(\psi)$ when ψ is a false value, as discussed in Evans and Shakhatreh (2008), namely, we calculate the prior probability that $RB(\psi) \geq k$ when $\theta \sim \Pi(\cdot \,|\, \psi_0), x \sim P_\theta$ and $\psi \sim \Pi_\Psi$ independently of (θ, x). So ψ is a false value in the generalized sense that it has no connection with the true value of the parameter and the data. We have the following result.

Theorem 10. The prior probability that $RB(\psi) \geq k$, when $\theta \sim \Pi(\cdot \,|\, \psi_0), x \sim P_\theta$ and $\psi \sim \Pi_\Psi$ independently of (θ, x), is bounded above by $1/k$.
Proof: We have that

$$\Pi(\cdot \,|\, \psi_0) \times P_\theta \times \Pi_\Psi \left(\frac{\pi_\Psi(\psi \,|\, T(x))}{\pi_\Psi(\psi)} \geq k \right) = M_T(\cdot \,|\, \psi_0) \times \Pi_\Psi \left(\frac{\pi_\Psi(\psi \,|\, t)}{\pi_\Psi(\psi)} \geq k \right)$$
$$= \int_{\mathcal{T}} \int_{\{\pi_\Psi(\psi \,|\, t)/\pi_\Psi(\psi) \geq k\}} \pi_\Psi(\psi) m_T(t \,|\, \psi_0) \, \nu_\Psi(d\psi) \, \mu_{\mathcal{T}}(dt)$$
$$\leq \frac{1}{k} \int_{\mathcal{T}} \int_{\{\pi_\Psi(\psi \,|\, t)/\pi_\Psi(\psi) \geq k\}} \pi_\Psi(\psi \,|\, t) m_T(t \,|\, \psi_0) \, \nu_\Psi(d\psi) \, \mu_{\mathcal{T}}(dt) \leq \frac{1}{k}.$$

So Theorem 10 says that it is *a priori* very unlikely that $RB(\psi)$ will be large when ψ is a false value, given that the true value is ψ_0. Putting $k = RB(\psi_0)$ gives the following result.

Corollary 11. $M_T(\cdot \,|\, \psi_0) \times \Pi_\Psi \left(m_T(t \,|\, \psi)/m_T(t) \geq RB(\psi_0) \right) \leq \left(RB(\psi_0) \right)^{-1}$.

So a large value of $RB(\psi_0)$ says there is little prior probability, when H_0 is true, of obtaining a larger value at another value of ψ and this can be interpreted as evidence of support for H_0.

Just as the results in Royall (2000) support the evidential use of likelihood ratios, Corollaries 9 and 11 support the evidential use of relative belief ratios and Bayes factors, where small values of these quantities are evidence against H_0 and large values are evidence in favor of H_0. But note that both of these results are *a priori* calculations given that H_0 holds. From fundamental considerations it is more appropriate that such assessments be *a posteriori* and moreover that alternatives to H_0 be considered. This is the role fulfilled by (6). Very small values of $RB(\psi_0)$ can be interpreted somewhat definitively as evidence against H_0 although, even in this case, we might want to see where most of the relative belief is being assigned via computing $\hat{\psi} = \arg\sup RB(\psi)$ to see if this represents a deviation from H_0 of practical significance. Large values of $RB(\psi_0)$ almost certainly need to be qualified by (6) and by looking at $(\hat{\psi}, RB(\hat{\psi}))$. It seems that the assessment of a hypothesis is a somewhat subtle process that involves more than the computation of a single number, whether that be a Bayes factor or a tail probability.

We stress that (6) should not be interpreted as a P-value. For example, suppose $RB(\psi_0) = 20$ and (6) equals 0.50. While $RB(\psi_0)$ suggests strong evidence in favour of H_0, the reliability of that evidence does not seem overwhelming. If (6) equaled 0.05, then the evidence in favour of H_0 doesn't seem at all reliable, while if (6) equaled 0.95, then the reliability of the evidence in favour of H_0 seems very strong. In all cases we recommend also reporting $(\hat{\psi}, RB(\hat{\psi}))$ as this is telling us where the data has increased belief the most, by what factor, and allows the issue of practical significance to be addressed.

While (6) is not a P-value it has some of the properties of a P-value when we consider its behaviour as a function of the data. For example, as we increase the amount of data then, under conditions, (6) converges to 0 when H_0 is false and converges to a uniformly distributed random variable when H_0 is true. Recall, however, that our interpretation of (6) in an application is *a posteriori,* namely, given the data. It is also worth remarking, as noted in Evans (1997), that, in the context of Example 4, Lindley's paradox is resolved as the limit of (6) as $\tau^2 \to \infty$ is equal to the P-value for the two-sided Z-text.

Vlachos and Gelfand (2003) and Garcia-Donato and Chen (2005) also propose a method for calibrating Bayes factors in the binary case, as discussed in Example 5, but their calibration involves tail probabilities based on the prior predictive distributions given by m_{H_0} and $m_{H_0^c}$. So their calibration is in the spirit of Corollary 9. The calibration based on (6) is, however, *a posteriori* and this is in accord with the axiom of conditional probability that says that all probability statements about unknown parameters should be conditional given the data.

5 Method of Concentration

We now consider the selection of a Ψ function when we are asked to assess $H_0 \subset \Theta$ that doesn't arise from a parameter of interest. In the Examples of Section 3 we gave some simple choices but such simplicity is not always available. We base the selection of a Ψ function for a general problem on the following idea. If H_0 is true, then we expect the observed data to lead to the posterior distribution of θ being more concentrated about H_0 than the prior distribution of θ. To measure the concentration of a distribution about H_0 we choose a measure of the distance $d_{H_0}(\theta)$ of θ from H_0, such that $d_{H_0}(\theta) = 0$ if and only if $\theta \in H_0$, and then see how closely the distribution of $d_{H_0}(\theta)$ concentrates about 0. Note that $\Psi(\theta) = d_{H_0}(\theta)$ generates H_0 via $H_0 = \Psi^{-1}\{0\}$. Often we can choose d_{H_0} so that $\Psi \in \mathcal{T}_{H_0}$ although we don't regard this as essential. The following example considers a natural choice.

Example 6. *Squared Euclidean distance.*

Suppose Θ is an open subset of R^k and let $d_{H_0}(\theta) = ||\theta - \theta_0(\theta)||^2$ where $\theta_0(\theta)$ is a point in H_0 that is closest to θ. For example, suppose H_0 is a linear subspace, namely, $H_0 = L(X)$ for some $X \in R^{k\times l}$ of rank l. Then $\theta_0(\theta) = X(X^tX)^{-1}X^t\theta$ and $\Psi(\theta) = ||\theta - \theta_0(\theta)||^2 = \theta^t(I - X(X^tX)^{-1}X^t)\theta$. Therefore, $d\Psi(\theta) = 2\theta^t(I - X(X^tX)^{-1}X^t)$ and $J_\Psi(\theta) = ||\theta - \theta_0(\theta)||^{-1}/\sqrt{2} = \Psi^{-1/2}(\theta)/\sqrt{2}$

so Ψ has constant volume distortion for each value of ψ. Obviously this generalizes to affine subspaces of R^k.

Now suppose $H_0 = \{\theta : \theta^t\theta = 1\}$, i.e., H_0 is the unit sphere and $\Psi(\theta) = ||\theta - \theta_0(\theta)||^2$. Then $\theta_0(\theta) = \theta/||\theta||$ and $\Psi(\theta) = ||\theta - \theta_0(\theta)||^2 = ||\theta||^2(1 - 1/||\theta||)^2 = (||\theta|| - 1)^2$ and so $d\Psi(\theta) = 4(||\theta|| - 1)\theta^t/||\theta||$ giving $J_\Psi(\theta) = 4|||\theta|| - 1| = 4\Psi^{-1/2}(\theta)$. Again Ψ has constant volume distortion at each value of ψ. We can easily generalize this to spheres with an arbitrary radius and center. Also we can generalize to ellipsoids by first reparameterizing so that the ellipsoid is a sphere in the new parameterization.

While there is nothing that forces us to choose d_{H_0} to be squared Euclidean distance, this choice often has some computational convenience and can be considered as a generalization of the use of variance in statistics as a measure of spread. As such it has intuitive appeal but we certainly do not rule out other choices. This reflects a certain arbitrariness in problems where we have not specified a parameter of interest, but we recall our discussion of Section 3 where we noted that consistency of prior beliefs requires that such a choice be made.

Given that we have chosen d_{H_0}, the question remains as to how we should compare the concentrations of the prior and posterior about H_0. In Evans, Gilula, and Guttman (1993) and Evans, Gilula, Guttman, and Swartz (1997) d_{H_0} was taken to be squared Euclidean distance and the prior and posterior distributions of Ψ were compared graphically. A more formal method would compute a Bayes factor and assess the reliability of the Bayes factor via (6) with $\Psi(\theta) = ||\theta - \theta_0(\theta)||^2$. There is a problem with this, however, as in Example 6 we have $J_\Psi(\theta) \equiv 0$ when $\theta \in H_0$ and as such we cannot use $\pi_\Psi(0 \,|\, T(x))/\pi_\Psi(0)$ to define $RB(0)$. The following example illustrates this and how to resolve the problem.

Example 7. *Squared Euclidean distance (continued).*

Suppose that Θ is an open subset of R^1 and $H_0 = \{\theta_0\}$. Now for each $\psi > 0$ we have that $\nu_{\Psi^{-1}\{\psi\}}$ is counting measure on $\Psi^{-1}\{\psi\} = \{\theta_0 - \psi^{1/2}, \theta_0 + \psi^{1/2}\}$. Therefore, $\pi_\Psi(\psi) = \{\pi(\theta_0 - \psi^{1/2}) + \pi_\Psi(\theta_0 + \psi^{1/2})\}\psi^{-1/2}$ and $\pi_\Psi(\psi \,|\, T(x)) = \{\pi(\theta_0 - \psi^{1/2} \,|\, T(x)) + \pi_\Psi(\theta_0 + \psi^{1/2} \,|\, T(x))\}\psi^{-1/2}$. Accordingly, we have that

$$RB(0) = \lim_{\psi \to 0} \pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi) \tag{8}$$

as the limit equals $\pi(\theta_0 \,|\, T(x))/\pi(\theta_0)$, namely, the Bayes factor in favor of H_0.

This result can be easily generalized to higher dimensions with $\nu_{\Psi^{-1}\{\psi\}}$ equal to surface area on the sphere of radius $\psi^{1/2}$. To see a specific case, suppose that we are sampling from a $N_k(\mu, I)$ distribution, with the prior given by $\mu \sim N_k(0, I)$, and we want to assess $H_0 : \mu = 0$ using the distance measure $\Psi(\mu) = \mu'\mu$. The prior distribution of Ψ is then chi-squared$(k, 0)$, and if we observe the value $\bar{x}$ based on a sample of n, the posterior distribution is $(n + 1)^{-1}$chi-squared$(k, n^2\bar{x}'\bar{x}/(n + 1))$. Note that both densities vanish at 0. As k increases the order of this zero increases because the surface area of a sphere of radius $\psi^{1/2}$ in R^k is proportional to $\psi^{(k-1)/2}$. In spite of this $RB(0) = (n + 1)^{k/2} \exp\{-n^2\bar{x}'\bar{x}/(n + 1)\}$ via (7). This shows that the need to evaluate

$RB(0)$ via (8) does not arise because the prior is assigning little prior probability to H_0 but rather is due to the geometry associated with Ψ.

In general we define $RB(0)$ by (8). We note that this agrees with the definition given previously when $\pi_\Psi(0) > 0$ and $\pi_\Psi(0 \,|\, T(x)) > 0$ but extends it to the case where both are 0.

A computational difficulty also arises when $\pi_\Psi(0) = 0$ and $\pi_\Psi(0 \,|\, T(x)) = 0$ as we will approximate $RB(0)$ by $RB(\psi_*)$ for ψ_* near 0 and this requires accurate estimates of $\pi_\Psi(\psi_*)$ and $\pi_\Psi(\psi_* \,|\, x)$. Since we will commonly use simulation methods, the fact that the prior and posterior densities vanish at 0 indicates there will not be many sampled values of ψ near 0 from either the prior or posterior. We note, however, that, whenever ψ_* is in the left tail of the prior, then $\Pi_\Psi\left(RB(\psi) \le RB(\psi_*) \,|\, T(x)\right)$ is a measure of the concentration of the posterior relative to the prior about H_0. So our approach is to choose ψ_* to be a left tail quantile of Π_Ψ that can be reliably estimated, e.g., the 0.05 quantile, and then use $\Pi_\Psi\left(RB(\psi) \le RB(\psi_*) \,|\, T(x)\right)$ to approximate (6). Actually, rather than relying on a single number, it makes sense to look at $\Pi_\Psi(RB(\psi) \le RB(\psi_*) \,|\, x)$ for several choices of ψ_* in the left tail of the prior, to determine how the posterior distribution of Ψ has concentrated about 0 relative to its prior distribution.

We use the following algorithm for the approximations. This requires that we have available exact or approximate samplers for both Π_Ψ and $\Pi_\Psi(\cdot \,|\, T(x))$. Let N be a positive integer and $\psi_{i/N}$ denote the i/N-th prior quantile of Ψ for $i = 0, \ldots, N$ where $\psi_0 = 0$ and $\psi_1 = \infty$. Also, let $\hat{F}$ denote an empirical prior cdf, $\hat{F}(\cdot \,|\, x)$ denote an empirical posterior cdf, $\psi_* = \psi_{i_0/N}$ and $i_0 \in \{1, \ldots, N\}$.

(1) Select M_1 and generate a sample $\psi_1, \ldots, \psi_{M_1}$ from Π_Ψ.

(2) Compute the estimates $\hat{\psi}_0, \hat{\psi}_{1/N}, \hat{\psi}_{2/N}, \ldots, \hat{\psi}_1$, using interpolation between sample quantiles, where $\hat{\psi}_0 = 0$ and $\hat{\psi}_1$ is the largest sample value.

(3) Select M_2 and generate a sample $\psi_{1,x}, \ldots, \psi_{M_2,x}$ from $\Pi_\Psi(\cdot \,|\, T(x))$.

(4) For $\psi \in [\hat{\psi}_{i/N}, \hat{\psi}_{(i+1)/N})$ estimate $RB(\psi)$ by

$$\hat{RB}(\psi) = \frac{\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x)}{\hat{F}(\hat{\psi}_{(i+1)/N}) - \hat{F}(\hat{\psi}_{i/N})} = N(\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x)).$$

(5) Estimate $\Pi_\Psi(RB(\psi) \le RB(\psi_{i_0/N}) \,|\, x)$ by the finite sum

$$\sum_{\{i:\hat{RB}(\hat{\psi}_{i/N}) \le \hat{RB}(\hat{\psi}_{i_0/N})\}} (\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x)). \tag{9}$$

The following result, which is proved in the Appendix, establishes the convergence of (9) to (6) as N, M_1, and M_2 grow.

Theorem 12. Suppose that $RB(\psi)$ is continuous in ψ and $RB(\psi)$ has a continuous posterior distribution. Then (9) converges almost surely to to $\Pi_\Psi(RB(\psi) \le RB(0) \,|\, x)$ as $N \to \infty, M_1 \to \infty$ and $M_2 \to \infty$.

6 Logistic Regression

We consider a Bayesian analysis of the logistic regression model. For simplicity we will discuss the case of a single predictor but note that adding more predictors is completely feasible. We develop a test for goodness of fit based on the method of concentration and show that this leads to a method for assigning a prior on model parameters that avoids problems with more typical choices.

Let $x_i \in R^1$ denote the value of a predictor $X, p_i = P(Y = 1 \,|\, X = x_i)$ and $\mu_i = \ln p_i/(1-p_i)$. Suppose first that X is a categorical predictor taking k values and we observe n_i observations when $X = x_i$. If we assume the observations are independent, which is primarily a matter of how we sampled, then the appropriate model is a k-fold product of binomial models. Depending on what we know *a priori*, we could select independent beta priors for the p_i, although other choices are certainly possible. For example, if we felt that we know virtually nothing about p_i, then we could take p_i to be uniformly distributed and note this implies that the prior distribution of μ_i is standard logistic via the transformation $p_i = \exp\{\mu_i\}/(1+\exp\{\mu_i\})$. In this case it is relatively simple to assign a prior to the p_i or μ_i and, in particular, assign one that is noninformative.

Suppose now that X is quantitative. In this situation it is common to relate the p_i to X via $\mu_i = \beta_1 + \beta_2 x_i$ or use a higher degree polynomial. Interest then is in making inference about the β_i and in particular assess $H_0 : \beta_2 = 0$. It is common to put something like $N(0, \sigma_i^2)$ priors on the β_i where the σ_i^2 are chosen large to reflect little information *a priori*. As is well-known, see for example Evans and Jang (2011a), there are problems with such a choice and this is illustrated in Figure 3 where we have plotted the prior density of p when $\beta_1, \beta_2 \sim N(0, \sigma^2)$, $x = 1$ and $\sigma = 20$. As σ grows all the prior probability for p piles up at 0 and 1 and so this is clearly a poor choice. For a general $(p_1, \ldots, p_k)$ it is not clear how we should choose a normal prior on (β_1, β_2) to reflect the information about the p_i.

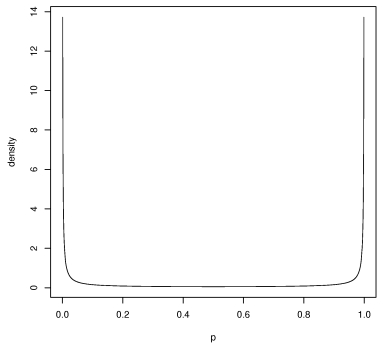

Figure 3: Prior density of of p when $\beta_1, \beta_2 \sim N(0, 20^2)$ and $x = 1$.

The strange behavior of such priors has been noted by other authors. Bedrick, Christensen and Johnson (1996, 1997), based on earlier work in Tsutukawa and Lin (1986), make the sensible recommendation that priors should instead be placed on the p_i, as these are the parameters for which we typically have prior

information. Their recommendation is that, when $\mu_i = \beta_1 + \beta_2 x_i$, two of the x_i values be selected and then Beta(α_i, β_i) priors be placed on the corresponding p_i. While this results in more sensible priors, it suffers from the need to select the two x_i values as the induced prior on the β_i will depend on this choice.

There is another issue that arises for the model with $\mu_i = \beta_1 + \beta_2 x_i$. Such a relationship holding for the k distinct observed values of X imposes a restriction on $(\mu_1, \ldots, \mu_k)$ that does not apply in the case when X is categorical. As such it is necessary to ask, before we make inferences about the β_i, if indeed such a relationship holds. Suppose then that we have no information about the relationship between Y and X other than that the y_i are conditionally independent given the x_i. In such a case it seems reasonable to assume that the p_i are i.i.d. $U(0,1)$. This implies that *a posteriori* $p_1, \ldots, p_k$ are independent with $p_i \sim \text{Beta}(1 + n_i \bar{y}_i, 1 + n_i(1 - \bar{y}_i))$, where $\bar{y}_i$ denotes the proportion of successes observed when $X = x_i$. If indeed we do have prior information about some or all of the p_i, then other beta distributions can be used for the priors.

We can now assess the logistic regression hypothesis using the method of concentration via generating from the prior and posterior distributions of the p_i and transforming to $\mu_i = \ln p_i/(1 - p_i)$. For if $\mu \in R^k$ is the vector of log-odds, then the logistic regression model holding is equivalent to $\mu \in H_0 = \mathcal{L}(V)$ where $V = (1_k, x), 1_k = (1, \ldots, 1)^t \in R^k$ and $x = (x_1, \ldots, x_k)^t$. Using Euclidean distance, which seems like a natural choice in this case, the method of concentration compares the prior and posterior distributions of $d_{H_0}(\mu) = ||\mu - V(V^tV)^{-1}V^t\mu||^2$. Consider an example.

Example 8. *Checking the logistic regression model.*

First we apply the goodness-of-fit test when the model is correct. For this consider the case $k = 3, x_1 = 0, x_2 = 1, x_3 = 2$ and $\beta_1 = 0.5, \beta_2 = -1$ so $p = (0.62, 0.38, 0.18)$. We generated data from the model, for varying common values of n_i. In Table 1 we present the generated values of the $n\bar{y}_i$, and the values of $\Pi_\Psi\left(RB(\psi) \leq RB(\psi_*) \mid T(x)\right)$ for various left-tail prior quantiles ψ_* of the prior together with the corresponding value of $RB(\psi_*)$. In Figure 4 we have plotted the posterior density of ψ and $RB(\psi)$ with the $0.05, 0.10, 0.20$ and 0.50 prior quantiles marked on the x-axis. We see that in each case the posterior distribution of the distance measure has concentrated in the left tail of the prior distribution of the distance measure and so we have no evidence against the model. We also consider an example where the simple logistic regression model is wrong. Let $k = 5$ with $x = (1, 3, 5, 7, 9)$ and $p = (0.875, 0.327, 0.107, 0.198, 0.908)$ so the relationship $\mu_i = \beta_1 + \beta_2 x_i$ does not hold for any choice of (β_1, β_2). In Table 2 we present the results for generated data from the model with these probabilities, and in Figure 5 we plot the prior and posterior densities of ψ and $RB(\psi)$. We see that there is clear evidence that the model does not hold.

This test of fit has low power when many of n_i are small. In design contexts we can select the values (x_i, n_i) to ensure sensitivity. In many situations, however, the data have many $n_i = 1$. In these circumstances we can typically group the x_i values and also fit higher order quadratic or cubic terms, and test for the higher order terms for the model checking. We illustrate this in Example 10.

Quantile	$\Pi_\Psi\left(RB(\psi) \leq RB(\psi_*) \mid T(x)\right)\ (RB(\psi_*))$		
	$n_i = 1$	$n_i = 5$	$n_i = 10$
	$(1, 0, 0)$	$(3, 2, 1)$	$(7, 4, 2)$
0.01	0.828 (1.059)	1.000 (2.073)	1.000 (2.714)
0.05	0.702 (1.048)	0.979 (2.067)	0.841 (2.611)
0.10	0.670 (1.046)	0.856 (2.011)	0.763 (2.567)

Table 1: Values of (8) and $RB(\psi_*)$ in Example 8 when ψ_* equals the 0.01, 0.05, and 0.10 prior quantiles and the model is correct.

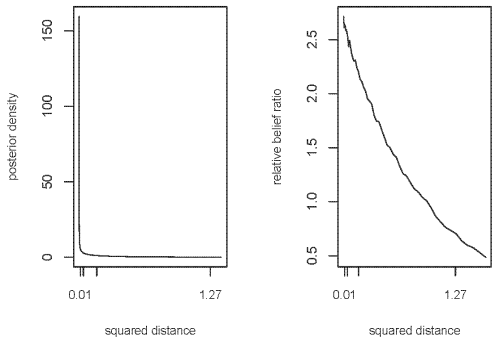

Figure 4: Plots of the posterior density of ψ and $RB(\psi)$ in Example 8 when the model is correct, $n_i = 10$, and with the 0.05, 0.10, 0.20 and 0.50 prior quantiles indicated.

Quantile	$\Pi_\Psi\left(RB(\psi) \leq RB(\psi_*) \mid T(x)\right)\ (RB(\psi_*))$		
	$n_i = 1$	$n_i = 5$	$n_i = 10$
	$(1, 1, 0, 0, 0)$	$(5, 3, 0, 2, 5)$	$(9, 2, 1, 2, 10)$
0.01	0.066 (0.761)	0.000 (0.017)	0.000 (0.000)
0.05	0.028 (0.732)	0.002 (0.046)	0.000 (0.001)
0.10	0.082 (0.793)	0.003 (0.060)	0.000 (0.003)

Table 2: Values of (8) and $RB(\psi_*)$ in Example 8 when ψ_* equals the 0.01, 0.05, and 0.10 prior quantiles and the model is wrong.

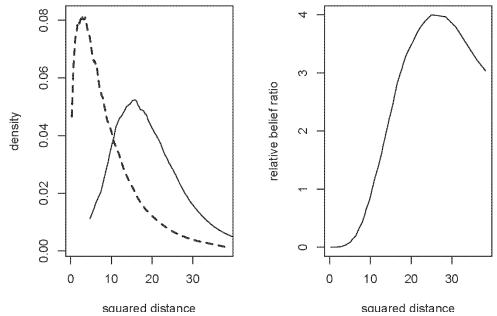

Figure 5: Plots of the prior (- -) and posterior densities (–) of ψ and $RB(\psi)$ in Example 8 when the model is wrong and $n_i = 10$.

Suppose that we obtain no evidence against the logistic regression model holding. Then appropriate inferences for the β_i are obtained using the conditional prior and posterior of μ given that $\mu \in H_0 = \mathcal{L}(V)$. Since $J_{d_{H_0}}(0) = 0$ we cannot use (3) to obtain the conditional prior density but we can proceed as follows. Let $V_\perp \in R^{k\times(k-2)}$ be such that its columns are an orthonormal basis of $\mathcal{L}^\perp(V)$. If $\psi = \Psi(\mu) = V_\perp'\mu$, then $J_\Psi(\mu) \equiv 1$ from which we conclude that $\Psi \in \mathcal{T}_{H_0}$ with $H_0 = \Psi^{-1}\{0\}$. Let Π denote the prior on μ with density π, and ν_{H_0} denote volume measure on H_0. Then by (3), $\pi(\mu)/\int_{H_0}\pi(\mu)\,\nu_{H_0}(d\mu)$ is the conditional prior density of μ given that $\Psi(\mu) = 0$. Now let $C_\epsilon = \{\mu : d_{H_0}(\mu) < \epsilon\}$ and and note that C_ϵ shrinks to H_0 as $\epsilon \downarrow 0$. Let $A \subset H_0$ and put $A_\epsilon = \{\mu : \mu_0(\mu) \in A\} \cap C_\epsilon$ so A_ϵ shrinks to A as $\epsilon \downarrow 0$. Then it is easy to show that $\Pi(A_\epsilon \,|\, C_\epsilon) \to \int_A \pi(\mu)\,\nu_{H_0}(d\mu)/\int_{H_0}\pi(\mu)\,\nu_{H_0}(d\mu)$ as $\epsilon \to 0$ and so $\pi(\mu)/\int_{H_0}\pi(\mu)\,\nu_{H_0}(d\mu)$ is the conditional prior density of μ given that $d_{H_0}(\mu) = 0$. Accordingly, when using a product of uniform priors on the p_i, the conditional prior density of μ, given H_0,is proportional to $\prod_{i=1}^k \exp\{\mu_i\}/(1+\exp\{\mu_i\})^2$. If we coordinatize H_0 by (β_1, β_2), then the conditional prior of (β_1, β_2) is proportional to $\prod_{i=1}^k \exp\{\beta_1 + \beta_2 x_i\}/(1+\exp\{\beta_1 + \beta_2 x_i\})^2$. A similar result holds for the conditional posterior. If we choose to put informative beta priors on some of the p_i, then it is still easy to determine the form of the conditional prior on (β_1, β_2).

Still it is somewhat difficult to see exactly what these priors are saying about the β_i and it is also difficult to see how to take into account the fact that, when two x_i values are close, then the p_i should be highly correlated. A well-known $N(0, d^2)$ approximation to the standard logistic distribution, as discussed in Camilli (1994), leads to conditional priors that are much easier to work with. The optimal choice of d, in the sense that it minimizes $\max_{x\in R^1}|\Phi(x/d) - e^x/(1+e^x)|$ is given by $d = 1.702$ and this leads to a maximum difference less than 0.01. Clearly this error will generally be irrelevant when considering priors for the probabilities in a logistic regression problem. So when μ is distributed $N(0, 1.702^2)$ we have that $e^\mu/(1+e^\mu)$ is approximately distributed $U(0,1)$ with the same maximum error. In Figure 6 we have plotted the density of $p = e^\mu/(1+e^\mu)$ when μ is $N(0, d^2)$ for various choices of d and we see that it is indeed approximately uniform when $d = 1.702$.

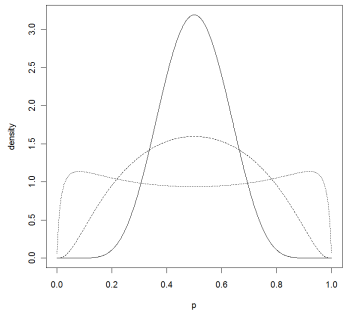

Figure 6. Plots of the density of $p = e^\mu/(1+e^\mu)$ when μ is $N(0, d^2)$ and $d = 0.5$ (–), $d = 1.0$ (- -), and $d - 1.702$ (...).

The prior $\prod_{i=1}^{k} \exp\{\mu_i\}/(1+\exp\{\mu_i\})^2$ induced on the logits by the uniform prior on $(p_1, \ldots, p_k)$ is then approximated by the $N_k(0, d^2 I)$ prior with $d = 1.702$. Now suppose that we condition on $\mu = V\beta$ where $V \in R^{k\times 2}$ and $V'V = I$, namely, V is column orthonormal. The conditional prior on β is proportional to $\exp\{-\beta' V' V\beta/2d^2\} = \exp\{-\beta'\beta/2d^2\}$, namely, the conditional prior on β is $N_2(0, d^2 I)$. Note that this says that the β_i are independent and approximately distributed standard logistic. Then, applying the standard logistic cdf F componentwise, we have that $p = F(u)$ where $u = V\beta \sim N_k(0, d^2 VV')$. Note that the $N_k(0, d^2 VV')$ distribution is singular but we can generate a value from it easily via $u = dVz$ where $z \sim N_2(0, I)$. From this we can simulate to obtain the conditional joint prior on p. We consider an example.

Example 9. *Equispaced points*

Suppose $x_i = a + (i-1)h$ for $i = 1, \ldots, k$. We then have that $v_{i1} = 1/\sqrt{k}$, $v_{i2} = 2\sqrt{3}\,(i - (k+1)/2)\,/\sqrt{k(k^2-1)}$ and

$$\begin{pmatrix} u_i \\ u_j \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, d^2 \begin{pmatrix} v_{i1}^2 + v_{i2}^2 & v_{i1}v_{j1} + v_{i2}v_{j2} \\ v_{i1}v_{j1} + v_{i2}v_{j2} & v_{j1}^2 + v_{j2}^2 \end{pmatrix}\right).$$

If $d = 1.702$ and the closer $v_{i1}^2 + v_{i2}^2$ is to 1 the closer the conditional prior distribution of p_i is to approximate uniformity and the closer $(v_{i1}v_{j1} + v_{i2}v_{j2})^2/(v_{i1}^2 + v_{i2}^2)(v_{j1}^2 + v_{j2}^2)$ is to 0 the less correlated are p_i and p_j. When $k = 2$, then $SD(u_1) = SD(u_2) = 1$ and $Corr(u_1, u_2) = 0$ and so $(p_1, p_2) \approx U(0,1)^2$. When $k = 3$, then $SD(u_1) = SD(u_3) = \sqrt{0.83}d = 1.55, Var(u_2) = \sqrt{0.33}d = .98, Corr(u_1, u_2) = Corr(u_2, u_3) = 0.63$, and $Corr(u_1, u_3) = -0.20$. When $k = 10$ then representative standard deviations are $SD(u_1) = \sqrt{0.35}d = 1.00$ and $SD(u_5) = \sqrt{0.10}d = 0.54$, while representative correlations are $Corr(u_1, u_2) = 0.99, Corr(u_1, u_7) = 0.07$ and $Corr(u_1, u_{10}) = -0.42$. When $k = 50$ the smallest standard deviation is $0.14d = 0.24$ and the largest is $0.28d = .48$ so the marginal conditional priors on the p_i have a reasonable degree of spread (see Figure 6). When $k = 100$ the smallest standard deviation is $0.10d = 0.17$ and the largest is $0.20d = 0.34$. So we see that variances are more extreme the farther we move away from the center and correlations are high and positive when the points are close together becoming negative when points are far apart. Given that F is a strictly increasing function these statements also apply to the p_i values.

Several questions remain to be addressed. The prior $\prod_{i=1}^{k} \exp\{\mu_i\}/(1+\exp\{\mu_i\})^2$ on the logits makes sense when we know absolutely nothing about the p_i. In situations where we feel we know something about p_i, e.g., we know that p_i is very small, it makes more sense to place a prior on p_i that reflects this. One could then choose location and scaling parameters for the logistic distribution to reflect to what our knowledge about p_i says about μ_i. Again we can approximate such a distribution by a normal distribution as it is easier to work with. Furthermore, given that we have available a predictor variable, it seems desirable that the prior we use reflect the fact that we know that when $|x_i - x_j|$ is small then $|p_i - p_j|$ is small too. This is accomplished by imposing a correlation between μ_i and μ_j that depends upon $|x_i - x_j|$. In the end we wind up choosing a $N_k(\alpha, \Sigma)$ prior on μ. Then the conditional prior on β is

proportional to $\exp\{-(V\beta-\alpha)'\Sigma^{-1}(V\beta-\alpha)/2d^2\}$ which is in turn proportional to $\exp\{-(\beta-(V'\Sigma^{-1}V)^{-1}V'\alpha)'V'\Sigma^{-1}V(\beta-(V'\Sigma^{-1}V)^{-1}V'\alpha)/2d^2\}$. Therefore, the conditional prior distribution of β is

$$N_2((V'\Sigma^{-1}V)^{-1}V'\alpha, d^2(V'\Sigma^{-1}V)^{-1}) \tag{10}$$

and $p = F(u)$ with

$$u = V\beta \sim N_k(V(V'\Sigma^{-1}V)^{-1}V'\alpha, d^2V(V'\Sigma^{-1}V)^{-1}V'). \tag{11}$$

So we choose (α_i, σ_{ii}) to reflect what we know about p_i through $p_i = F((z-\alpha_i)/d\sqrt{\sigma_{ii}})$ where $z \sim N(0,1)$. For example, if we think that p_i is in the interval $(0.1, 0.2)$ with prior probability 0.9 then $0.9 = \Phi(\alpha_i + d\sqrt{\sigma_{ii}}F^{-1}(0.2)) - \Phi(\alpha_i + d\sqrt{\sigma_{ii}}F^{-1}(0.1))$ and placing another prior probability restriction on p_i will allow us to solve for α_i and σ_{ii}. Choosing $(\alpha_i, \sigma_{ii}) = (0, d^2)$ corresponds to being noninformative about p_i. The issue of correlations is more difficult but a plausible approach here is to take $Corr(\mu_i, \mu_j) = \exp\{-(x_i - x_j)^2/2l^2\}$ where $l \geq 0$ is a hyperparameter chosen to reflect how quickly we believe p_i and p_j will become alike as $|x_i - x_j| \to 0$. Note that if we take $l = 0$, then the p_i are *a priori* uncorrelated and when $l = \infty$, then the p_i are completely dependent. The choice of l reflects what we know about the dependence of the probabilities on the predictor. Perhaps a natural approach to choosing l is to select (μ_i, μ_j) and then choose l so that $Corr(\mu_i, \mu_j)$ equals some specific value. Alternatively, we could consider placing a prior on l. Note that, when $x_i \to x_j$, then the $N_k(\alpha, \Sigma)$ prior converges to the $N_{k-1}(\alpha_{-i}, \Sigma_{-i})$ distribution where $(\alpha_{-i}, \Sigma_{-i})$ is obtained from (α, Σ) by deleting all entries associated with the i-th coordinate. So there is a coherency among priors between the situations where we consider μ_i and μ_j as possibly very different, because $|x_i - x_j|$ is large, and when we think of them as virtually identical because $|x_i - x_j|$ is quite small. Of course 'large' and 'small' are application dependent as this depends on the meaning of the predictor.

We now consider a real data application of this to a quadratic logistic regression model.

Example 10. *Gender-Height data.*

We consider predicting gender from height (Ht) measurements in cm from a data set on 102 male and 100 female athletes collected by the Australian Institute of Sport. There are 147 distinct values of Ht and we grouped these to form 21 cells each of length 3 cm ranging from 147 cm to 211 cm. The value of the predictor variable was taken to be the midpoint of each interval. This was done to reduce the size of the correlation matrix and it is felt that little was lost in describing the relationship. The raw data can be found in Sheather (2008).

We considered the model $\text{logit}(p) = \beta_1 + \beta_2 Ht + \beta_3(Ht)^2$. For the prior on the 21 distinct μ_i we chose a $N_{21}(0, d^2R)$ distribution with $l = 2$ so that the correlation between the μ_i in adjacent cells equals 0.32. In Figure 7 we have plotted the prior density, posterior density and the relative belief ratio of the squared distance for the goodness of fit test. We can see from this that there is evidence in favour of the logistic model. In fact the goodness-of-fit test for the

quadratic model gave the values for $(\Pi_\Psi(RB(\psi) \leq RB(\psi_*) \mid T(x)), RB(\psi_*))$ as $(1.000, 5.701), (0.829, 2.956)$ and $(0.674, 2.159)$ when ψ_* is the $0.01, 0.05$ and 0.10 prior quantile, respectively.

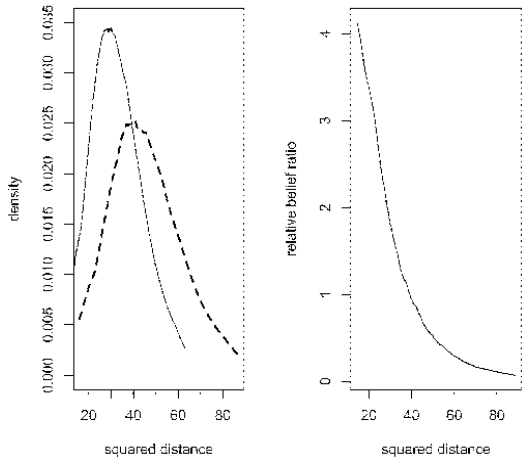

Figure 7: Plots of the prior density, posterior density and relative belief ratio of the squared distance in Example 10.

In Figure 8 we have plotted the marginal priors and posteriors of the β_i.

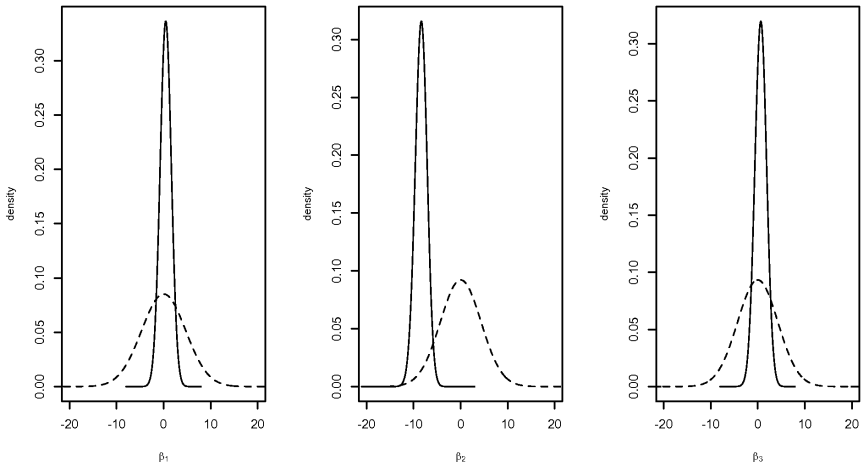

Figure 8: Plots of the marginal prior (- -) and posterior (–) densities of the β_i coefficients in Example 10.

For the hypothesis $H_0 : \beta_3 = 0$ we obtained $RB(0) = 2.99$ with (6) equal to 0.59 and $(\hat{\beta}_3, RB(\hat{\beta}_3)) = (0.72, 3.47)$ so this is evidence in favour of $\beta_3 = 0$. For

the hypothesis $H_0 : \beta_2 = 0$ we obtained $RB(0) = 1.6 \times 10^{-11}$ which we know by Theorem 4 immediately provides strong evidence against H_0.

7 Conclusions

We have shown that, when a hypothesis H_0 has 0 prior probability with respect to a prior on Θ, a Bayes factor and a relative belief ratio of H_0 can be sensibly defined without the need to introduce a discrete mass on H_0. Furthermore, we have developed a resolution of an inconsistency in Bayesian analyses that arises when a prior is used within H_0 that is not induced from the prior on the whole parameter space. When H_0 is generated by a parameter of interest, this inconsistency is easily avoided by using the conditional prior given H_0. When a problem is ill-specified, namely, we don't specify a parameter of interest, then we need to select one that generates H_0 to avoid this inconsistency. A natural approach in this situation is use a distance measure from H_0 as the parameter of interest, as this is seen to be equivalent to comparing the concentration of the posterior about H_0 with the concentration of the prior about H_0. This was applied to obtaining a Bayesian goodness of fit test for the logistic regression model and in turn this was used to obtain a prior on the regression coefficients that was induced from a prior on the success probabilities. This approach is seen to avoid problems associated with attempting to place priors directly on regression coefficients.

In general, we have argued that computing a Bayes factor, a measure of the reliability of the Bayes factor via a posterior tail probability, and the point where the Bayes factor is maximized together with its Bayes factor, provides a logical, consistent approach to hypothesis assessment. Various inequalities were derived that support the use of the Bayes factor in assessing either evidence in favour or against a hypothesis. The capacity to provide evidence either for or against a hypothesis is a significant advantage of the Bayesian approach to hypothesis assessment.

We note further that Evans and Shakhatreh (2008) establishes various optimal properties for credible regions and associated tests derived from (6). Evans and Jang (2011b) proves that $\hat{\psi}$, the point maximizing $RB(\psi)$, has optimal decision-theoretic properties, namely, it is either a Bayes rule or a limit of Bayes rules.

Appendix

Proof of Theorem 12: If $C = \{\psi : RB(\psi) = RB(0)\}$, then $\Pi_\Psi(C \,|\, x) = 0$ and so $\Pi_\Psi(RB(\psi) \leq RB(0) \,|\, x) = \Pi_\Psi(RB(\psi) < RB(0) \,|\, x)$. We use the following result.

Lemma (i) $\Pi_\Psi(\{\psi : RB(\psi) < RB(0)\} \triangle \lim(A_N \backslash C) \,|\, x)) = 0$ where $A_N = \cup\{(\psi_{i/N}, \psi_{(i+1)/N}] : N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x)) < NF(\psi_{1/N} \,|\, x)\}$,

(ii) $\Pi_\Psi(\limsup B_N \,|\, x) = 0$ where $B_N = \cup\{(\psi_{i/N}, \psi_{(i+1)/N}] : N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x)) = NF(\psi_{1/N} \,|\, x)\}$.

Proof: (i) Suppose $RB(\psi) < RB(0)$ and let $\delta = RB(0) - RB(\psi)$. Since $RB(0) = \lim_{N\to\infty} F(\psi_{1/N} \,|\, x)/F(\psi_{1/N}) = \lim_{N\to\infty} NF(\psi_{1/N} \,|\, x)$ there exists N_ψ such that for all $N > N_\psi$ we have $NF(\psi_{1/N} \,|\, x) > RB(0) - \delta/2$. For each N there exist unique prior quantiles such that $\psi_{i/N}(\psi) < \psi \leq \psi_{(i+1)/N}(\psi)$. By the convergence of these quantiles to ψ as N increases, which implies the convergence of $N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x))$ to $RB(\psi)$, there exists $N'_\psi > N_\psi$ such that for all $N > N'_\psi$ we have that $N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x)) < RB(0) - \delta/2$. This proves that $\{\psi : RB(\psi) < RB(0)\} \subset \liminf A_N$. Suppose now that $\psi \in \limsup A_N$. Then there are infinitely many values of N such that $N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x)) < NF(\psi_{1/N} \,|\, x)$. This implies that $RB(\psi) \leq RB(0)$ and so $\limsup(A_N \backslash C) = (\limsup A_N) \backslash C \subset \{\psi : RB(\psi) < RB(0)\} \subset (\liminf A_N) \backslash C = \liminf(A_N \backslash C)$ and we are done.

(ii) Suppose that $\psi \in \limsup B_N$. Then $(\psi_{i/N}(\psi), \psi_{(i+1)/N}(\psi)] \subset B_N$ for infinitely many N and this implies that $\psi \in C$ and the result follows.

Now let $\epsilon > 0$ and suppose $i_0 = 1$. From part (i) of the Lemma we have that $\Pi_\Psi(RB(\psi) \leq RB(0) \,|\, x) = \lim \Pi_\Psi(A_N \backslash C \,|\, x) = \lim \Pi_\Psi(A_N \,|\, x)$ so there exists N_0 such that $|\Pi_\Psi(RB(\psi) \leq RB(0) \,|\, x) - \Pi_\Psi(A_N \,|\, x)| < \epsilon/3$ for all $N > N_0$. By part (ii) of the Lemma there exits N_{00} such that for all $N > N_{00}$ we have that $\Pi_\Psi(B_N \,|\, x) < \epsilon/6$. Hereafter we suppose that $N > \max\{N_0, N_{00}\}$.

We now prove that for all M_1 and M_2 large enough, (1) is within ϵ of $\Pi_\Psi(RB(\psi) \leq RB(0) \,|\, x)$. Let $S_N = \{i : N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x)) < NF(\psi_{1/N} \,|\, x)\}$ and note that $\Pi_\Psi(A_N \,|\, x) = \sum_{i\in S_N}(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x))$. By Theorem 2.3.1 of Serfling (1980) we have that $\hat{\psi}_{i/N} \to \psi_{i/N}$ almost surely as $M_1 \to \infty$ and $M_2 \to \infty$. Also, by the Glivenko-Cantelli theorem, we have that $\hat{F}(\psi)$ converges almost surely and uniformly to $F(\psi)$ and similarly $\hat{F}(\psi \,|\, x)$ converges to $F(\psi \,|\, x)$ as $M_1 \to \infty$ and $M_2 \to \infty$. In particular, this implies that $\overset{\wedge}{RB}(0)$ converges almost surely to $NF(\psi_{1/N} \,|\, x)$.

Suppose $i \in S_N$. Let $\delta = NF(\psi_{1/N} \,|\, x) - N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x))$. Then, for all M_1, M_2 large enough, $\overset{\wedge}{RB}(0) > NF(\psi_{1/N} \,|\, x) - \delta/2$ and

$$N(\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x)) < N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x)) + \delta/2.$$

This implies that $i \in \hat{S}_N = \{i : \overset{\wedge}{RB}(\hat{\psi}_{i/N}) \leq \overset{\wedge}{RB}(0)\}$ for all M_1, M_2 large enough and so $S_N \subset \hat{S}_N$ for all M_1, M_2 large enough. Furthermore, the contribution $(\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x))$ that this index makes to the sum (1) converges to $(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x))$ as $M_1 \to \infty$ and $M_2 \to \infty$.

If $i \notin S_N$ and $N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x)) > NF(\psi_{1/N} \,|\, x)$, then the same argument shows that $i \notin \hat{S}_N$ for all M_1, M_2 large enough. Since $\#(\hat{S}_N) \leq N$, then for for all M_1, M_2 large enough, we have that $\sum_{i\in\hat{S}_N}(\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x)) = \sum_{i\in S_N}(\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x)) + \sum_{i\in\hat{S}_N\backslash S_N}(\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x))$ where the second sum contains at most those terms corresponding

to i where $N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{(i+1)/N} \,|\, x)) = NF(\psi_{1/N} \,|\, x)$. The first sum converges almost surely to $\Pi_\Psi(A_N \,|\, x)$ and the limit supremum of the second term is bounded above by $\sum_{D(N,x)}(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x)) = \Pi_\Psi(B_N \,|\, x)$ where $D(N,x) = \{i : N(F(\psi_{(i+1)/N} \,|\, x) - F(\psi_{i/N} \,|\, x)) = NF(\psi_{1/N} \,|\, x)\}$. So for all M_1, M_2 large enough we have $|\sum_{i \in S_N}(\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x)) - \Pi_\Psi(A_N \,|\, x)| < \epsilon/3$ and $\left|\sum_{i \in \hat{S}_N \setminus S_N}(\hat{F}(\hat{\psi}_{(i+1)/N} \,|\, x) - \hat{F}(\hat{\psi}_{i/N} \,|\, x))\right| < \epsilon/3$ and this finishes the proof. The proof of the more general case, where i_0 is not constrained to be 1, follows easily by noting that $\psi_{i_0/N} \to 0$ as $N \to \infty$.

References

Bedrick, E.J., Christensen, R., and Johnson, W. (1996) A new perspective on priors for generalized linear models. Journal of the American Statistical Association, 91, 436, 1450-1460.

Bedrick, E.J., Christensen, R., and Johnson, W. (1997) Bayesian binomial regression: predicting survival at a trauma center. The American Statistician, 51, 3, 211-218.

Berger, J.O. and Delampady, M. (1987) Testing Precise Hypotheses. Statistical Science, 2, 317-335.

Berger, J.O., Liseo, B., and Wolpert, R.L. (1999) Integrated likelihood methods for eliminating nuisance parameters. Statistical Science, 14, 1, 1-28.

Berger, J.O. and Perrichi, R.L. (1996). The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association 91:10-122.

Camilli, G. (1994) Origin of the scaling constant d = 1.7 in item response theory. Journal of Educational and Behavioral Statistics, Vol. 19, No. 3, 293-295.

Dickey, J.M. and Lientz, B.P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. Annals of Mathematical Statistics, 41, 1, 214-226.

Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. Annals of Statistics, 42, 204-223.

Evans, M., Gilula, Z., and Guttman, I. (1993) Computational issues in the Bayesian analysis of categorical data : loglinear and Goodman's RC model. Statistica Sinica, 3, 391-406.

Evans, M., Gilula, Z., Guttman, I., and Swartz, T. (1997) Bayesian analysis of stochastically ordered distributions of categorical variables. Journal of the American Statistical Association, 92, 437, 208-214.

Evans, M. and Jang, G-H. (2011a) Weak informativity and the information in one prior relative to another. To appear in Statistical Science.

Evans, M. and Jang, G-H. (2011b) Inferences from prior-based loss functions. Technical Report No. 1104, Dept. of Statistics, U. of Toronto.

Evans, M. and Shakhatreh, M. (2008) Optimal properties of some Bayesian inferences. Electronic Journal of Statistics, 2, 1268-1280.

Garcia-Donato, G. and Chen, M-H. (2005) Calibrating Bayes factor under prior predictive distributions. Statistica Sinica, 15, 359-380.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) Bayesian Data Analysis, Second Edition. Chapman and Hall/CRC, Boca Raton, FL.

Jeffreys, H. (1935) Some Tests of Significance, Treated by the Theory of Probability. Proceedings of the Cambridge Philosophy Society, 31, 203- 222.

Jeffreys, H. (1961) Theory of Probability (3rd ed.), Oxford, U.K.: Oxford University Press.

Kass, R.E. and Raftery, A.E. (1995) Bayes factors. Journal of the American Statistical Association, 90, 430, 773-795.

Lavine, M. and Schervish, M.J. (1999) Bayes Factors: what they are and what they are not. The American Statistician, 53, 2, 119-122.

O'Hagan, A. (1995) Fractional Bayes factors for model comparisons (with discussion). Journal of the Royal Statistical Society B 56:3-48.

Robert, C.P., Chopin, N. and Rousseau, J. (2009) Harold Jeffreys's Theory of Probability Revisited (with discussion). Statistical Science, 24, 2, 141-172.

Royall, R. (1997) Statistical Evidence. A likelihood paradigm. Chapman and Hall, London.

Royall, R. (2000) On the probability of observing misleading statistical evidence (with discussion). Journal of the American Statistical Association, 95, 451, 760-780.

Rudin, W. (1974) Real and Complex Analysis, Second Edition. McGraw-Hill, New York.

Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. John Wiley and Sons, New York.

Sheather, S.J. (2008) A Modern Approach to Regression with R. Springer, New York.

Tjur, T. (1974) Conditional Probability Models. Institute of Mathematical Statistics, University of Copenhagen, Copenhagen.

Tsutukawa, R.K. and Lin, H.Y. (1986) Bayesian estimation of item response curves. Psychometrika, 51, 251-267.

Verdinelli, I. and Wasserman, L. (1995) Computing Bayes factors using a generalization of the Savage-Dickey density ratio. Journal of the American Statistical Association, 90, 430, 614-618.

Vlachos, P.K. and Gelfand, A.E. (2003) On the calibration of Bayesian model choice criteria. Journal of Statistical Planning and Inference, 111, 223-234.